

MoDIL: Detecting INDEL Variation with Mixtures of Distributions

Seunghak Lee¹, Fereydoon Hormozdiari², Can Alkan³, Michael Brudno^{1,4}

1 Introduction

Human genetic variation comes in a wide range of sizes - from SNPs and very small (single nucleotide) insertions/deletions (indels) to large-scale “structural” variations, where kilo base pairs of the genome are inserted, deleted, inverted, or duplicated. Several methods for the identification of both small scale variants (SNPs and insertions/deletions $< 10\text{bp}$) [2, 3, 8] and large scale ones (inversions and large insertions/deletions) [10, 6, 7, 5, 4] have been developed, and their discovery and cataloguing is well underway. Simultaneously, one would expect there should also be a large amount of “medium-sized” variation: insertions and deletions of 10 to 50 nucleotides. Currently, however, there are no methods, either computational or wet lab, for high-throughput detection of these medium sized polymorphisms. In this work we develop a method to find these variants (10–50bp) by relying on the high clone coverage of many NGS datasets. While deviation by several of these clones from the expected insert size is to be expected, a deviation by a very large number (even by a small amount) would be indicative of an insertion or deletion. Here, we show a rigorous method for identifying and analyzing indels, especially medium sized indels using this intuition.

2 Algorithm for Indel Detection

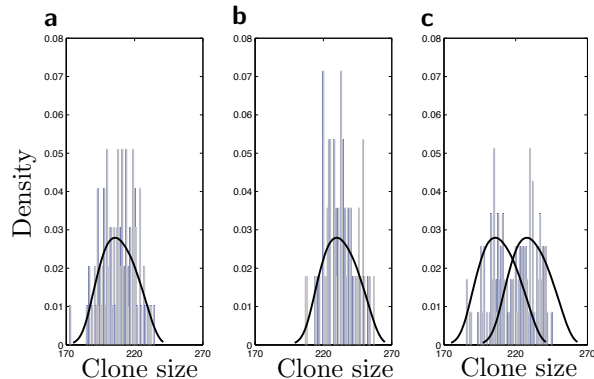


Figure 1: Observed distribution of mapped distances within a cluster for cases of (a) no indel, (b) homozygous deletion and (c) heterozygous deletion, superimposed on distribution of insert sizes, $p(Y)$ (See text for details).

Our algorithm is given a list of mapping of all matepairs to the reference genome. Each matepair consists of two reads, and the distance between them is referred to as the mapped distance. Unlike previous methods [10, 7, 4], which identify structural variations based on a pre-specified number of matepairs with a mapped distance significantly different from the insert size, our algorithm operates on the whole distribution of the observed mapped distances, rather than the outliers. Fig. 1 illustrates the essence of our method. We take advantage of the distribution of insert sizes in the sequenced library (we call this distribution $p(Y)$, and estimate it from all of the mapped distances, genome wide). Given a particular genomic location i , we define the corresponding cluster C_i as all of the matepairs that overlap the location (the left read of the pair is to the left of the position, and the right read is to the right). We will call the distribution of the observed mapped distances in a cluster $p(C_i)$. A cluster that comes from a location that has no indel polymorphisms will have mapped distances that follow the same distribution as the size of the clones in the library ($p(C_i)$ and $p(Y)$ will be identically distributed). This is illustrated (for our dataset) in Fig. 1a. If, on the other hand, there has been a homozygous insertion or deletion at this location, the distribution $p(C_i)$ will shift (Fig. 1b). In case of

¹Department of Computer Science, University of Toronto, Canada.

²School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.

³Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, WA, USA.

⁴Banting and Best Dept. of Medical Research, University of Toronto, Canada.

heterozygous indel, we will observe that $p(C_i)$ consists of two distributions, shifted and non-shifted $p(Y)$ s (Fig. 1c). Furthermore, the size of the indel event can be estimated with high confidence; its expected size follows a Gaussian distribution with mean $\mu = \mu_{p(y)} - \mu_{p(C_i)}$ and $\sigma = \sigma_{p(y)}/\sqrt{n}$, where n is the number of matepairs in the cluster. Note that expected size of indel is normally distributed regardless of observed distribution of mapped distances according to the central limit theorem. As the number of matepairs in each cluster grows, our confidence in the size of the indel will increase, allowing for the prediction of progressively smaller indels with higher coverage.

In order to estimate indel sizes, we model the random variable of the expected size of indel with two random variables, one for each haplotype. We call this probabilistic model *Mixture of Distributions* (MoD) since the observed distribution will be a mixture of distributions of two random variables. Given a cluster, we identify the two distributions which have the fixed shape of $p(Y)$ and arbitrary means that best fit the observed data using the Kolmogorov-Smirnov goodness of fit test [9]. The means of the two distributions are found using the Expectation-Maximization algorithm, while appropriate Bayesian priors are used to prevent over-fitting.

3 Results

Table 1: Overlap between the predictions of the MoDIL algorithm and the short indels discovered by Kidd et al. and Mills et al. The fraction of Kidd et al. indels present in our results indicates a low False Negative Rate (FNR) for our algorithm for indels ≥ 20 bp, and lower sensitivity for shorter indels. The large amount of overlap between our results and the Mills et al. data over all indel sizes indicates a strong correlation between them.

		MoDIL	Kidd et al.			Mills et al.		
Length	Type	Total	Total	Found	FNR	Total	Overlapping	% Overlap
≥ 20 bp	INS	1,607	101	91	0.10	6,240	260	0.16
	DEL	3,646	244	231	0.05	10,742	701	0.19
15-19bp	INS	592	124	57	0.54	3,096	115	0.19
	DEL	2,584	183	109	0.40	3,698	162	0.06
10-14bp	INS	257	373	56	0.85	8,615	125	0.49
	DEL	1,157	601	171	0.72	9,990	237	0.20

We have applied our model to the whole genome shotgun reads generated by Illumina for the Yoruban HapMap individual NA18507 [1]. This data provided 40x read coverage and 120x clone coverage for the NCBI reference genome build 35. The dataset was mapped to the reference NCBI human genome with mrFAST alignment tool (Alkan & Hormozdiari, unpublished) with 2-edit distance allowed. The resulting dataset had 292,190,651 templates, giving us 20x clone coverage of the genome. We required each cluster to have at least 20 matepairs to minimize false positives. Using our approach we discovered 2,529 insertions 7,716 deletions in the Yoruban individual genome relative to the NCBI reference genome (False Discovery Rate < 0.0001). These ranged in size from 5 to 119 nucleotides for insertions, and 5 to 334,646 nucleotides for deletions. In validation of our results, as shown in Table 1, we observed significant overlap between our indel calls and the short indels discovered by Kidd et al. with the same individual and Mills et al. with 36 human genomes. Our experimental results demonstrate that MoDIL can identify, with high sensitivity, indels ≥ 20 bp, while accurately estimating the true size of the variants. The size correlation of overlapping indels (20-500bp) between Mills et al. results and our indel calls was very strong ($r^2 = 0.99$).

References

- [1] D. R. Bentley et al. *Nature*, 456(7218):53–59, November 2008.
- [2] K. Chen et al. *Genome Res*, 17(5):659–666, May 2007.
- [3] L. W. Hillier et al. *Nature Methods*, 5(2):183, 2008.
- [4] F. Hormozdiari, C. Alkan, E. E. Eichler, and S. Sahinalp. *RECOMB 2009 in press*, 2009.
- [5] J. M. Kidd et al. *Nature*, 453(7191):56–64, 2008.
- [6] J. O. Korbel et al. *Science*, 318(5849):420, 2007.
- [7] S. Lee, E. Cheran, and M. Brudno. *Bioinformatics*, 24(13):i59–i67, 2008.
- [8] H. Li, J. Ruan, and R. Durbin. *Genome Research*, 18(11):1851, 2008.
- [9] F.J. Massey. *Journal of the American Statistical Association*, 46(253):68–78, 1951.
- [10] E. Tuzun et al. *Nature Genetics*, 37:727–732, 2005.